Verification of Eta–RSM Short-Range Ensemble Forecasts

THOMAS M. HAMILL AND STEPHEN J. COLUCCI

Department of Soil, Crop, and Atmospheric Sciences, Cornell University, Ithaca, New York

(Manuscript received 9 May 1996, in final form 8 August 1996)

ABSTRACT

Motivated by the success of ensemble forecasting at the medium range, the performance of a prototype short-range ensemble forecast system is examined. The ensemble dataset consists of 15 case days from September 1995 through January 1996. There are 15 members of the ensemble, 10 from an 80-km version of the eta model and five from the regional spectral model. Initial conditions include various in-house analyses available at the National Centers for Environmental Prediction as well as bred initial conditions interpolated from the medium-range forecast ensemble. Forecasts from the 29-km mesoeta model were archived as well for comparison.

The performance of the ensemble is first evaluated by the criterion of ''uniformity of verification rank.'' Assuming a perfect forecast model, equally plausible initial conditions, and the verification is a plausible member of the ensemble, these imply the verification when pooled with the 15 ensemble forecasts and sorted is equally likely to occur in each of the 16 ranks. Hence, over many independent samples, a histogram of the rank distribution should be nearly uniform. Using data from the ensemble forecasts, rank distributions were populated and found to be nonuniform. This was determined to be largely a result of model and initial condition deficiencies and not problems with the verification data. The uniformity of rank distributions varied with atmospheric baroclinicity for midtropospheric forecast variables but not for precipitation forecasts.

Examination of the error characteristics of individual ensemble members showed that the assumption of identical errors for each member is not met with this particular ensemble configuration, primarily because of the use of both bred and nonbred initial conditions in this test. Further, there were both differences in the accuracy of eta and regional spectral model bred member forecasts.

The performance of various summary forecasts from the ensemble such as its mean showed that the ensemble can generate forecasts that have similar or lower error than forecasts from the 29-km mesoeta, which was approximately equivalent in computational expense. Also, by combining the ensemble forecasts with rank information from other cases, reliable ensemble precipitation forecasts could be created, indicating the potential for useful probabilistic forecasts of quantitative precipitation from the ensemble.

1. Introduction

The use of ensemble methodologies has resulted in dramatic improvements in the skill of medium-range weather forecasts (Tracton and Kalnay 1993; Toth and Kalnay 1993; Molteni et al. 1996). Motivated by this success, research has begun in the application of ensemble methodologies to short-range forecasts (0–48 h).

Ensemble techniques have been adopted as a practical method for numerical weather prediction given the atmosphere's sensitive dependence on the initial condition (Lorenz 1963). Small errors in the initial condition (IC) grow exponentially during the forecast integration, so a single deterministic forecast will eventually be useless as guidance. Ensemble forecasting (Epstein 1969; Leith 1974) adopts the alternative goal of predicting the probability of future weather conditions. Here, a varied set of ICs are generated that are all consistent with the observations and their errors. Separate deterministic forecasts are integrated from each IC, and the relative frequency of weather outcomes are used to estimate a forecast probability distribution.

At 1–2 weeks lead time, even the planetary-scale flow shows the effects of sensitive dependence on the IC, and hence ensemble methodologies have proved beneficial for medium-range forecasting of planetary wave patterns. For short-range forecasts, the focus is on improving the predictions of specific weather elements, such as cloud cover, precipitation amount, and temperature. Through 48 h, the planetary scale is relatively predictable; however, synoptic and subsynoptic features are less predictable (Lorenz 1969; Livingston and Shaefer 1990). If the weather elements of interest are affected by these smaller-scale, more chaotic features, as seems plausible, ensemble methodologies may prove similarly beneficial for the short range.

The optimal use of available computer power is a contentious issue. Until only the last few years, it has been assumed in short-range weather prediction that the most beneficial use of newly available computer re-

Corresponding author address: Thomas M. Hamill, Department of Soil, Crop, and Atmospheric Sciences, 1126 Bradfield Hall, Cornell University, Ithaca, NY 14853.
E-mail: tmh8@cornell.edu

sources is for the execution of weather forecast models on finer grid meshes and using more complex model physics (Brooks et al. 1992; Brooks and Doswell 1993; Harrison 1994). Running a deterministic forecast at higher resolution allows more scale interaction, reduces the errors due to finite-difference approximations, permits more realistic treatment of cloud-scale processes, and improves the forecasts of blocks (Tracton 1990). There is a vast literature of the improvements achieved with higher resolution models. However, past improvements from finer resolution do not guarantee continuation of the trend. In the United States, between 1981 and 1996, the state-of-the art forecast model has increased in resolution from approximately 190 km (Newell and Deaven 1981) to 29 km (Black 1994; Rogers et al. 1996), a resolution increase of 6.5 times. McPherson (1991) suggests computational power will be available early in the next century to run limited-area forecast models at 5-km resolution. While there will undoubtedly be improvements to the forecast from increased resolution, the main desire is the ability to accurately forecast mesoscale detail from such high-resolution models. The evidence for this is more mixed. For example, Reynolds et al. (1994) found that most forecast errors in the midlatitudes are now attributable to predictability error growth rather than model deficiencies; that is, problems with the IC. Similarly, experiments at the European Centre for Medium-Range Weather Forecasts (ECMWF) (Simmons et al. 1995) have shown that forecasts at T213 resolution amplify initial errors more quickly than previous, lower-resolution versions of the model, and Kuo and Reed (1988) did not find significant improvement of explosive cyclogenesis from increased resolution. Without consistent and reliable mesoscale information in the IC, the predictability of mesoscale features in the forecast is suspect unless the feature evolved due to terrain or large-scale forcing. While these no doubt frequently happen, there are many forecast situations where mesoscale features are organized by mesoscale detail not captured by current observing systems, and hence the forecast mesoscale generated will not be trustworthy. The ensemble approach to forecasting provides a way of addressing the uncertainty in the IC. Assuming a halving of model resolution decreases CPU usage sixteenfold, for the same computational time as a 5-km, single-integration forecast, an 8-member ensemble forecast could be run at 8.4 km, a 16-member ensemble 10-km resolution, or a 256-member at 20 km. Similarly, further into the future, the choice may be between 0.5-km single integrations or 1-km ensembles. Hence, even if short-range ensemble forecasting proves not to be beneficial given today's coarser grid structure, the exponential increase in computer power indicates its eventual relevance.

Because of the potential relevance now and its future relevance, ensemble methodologies applied to the short range are being actively explored (Mullen and Baumhefner 1994; Manikin 1995; Brooks et al. 1995; Brooks et al. 1996; Hamill and Colucci 1996). Since the experience with ensemble forecasts specifically for the short range is scant, questions are currently more numerous than answers. What is the best method for generating ICs? Should an ensemble consist only of perturbations to the ICs, or include ''perturbations'' to model physics such as different convective trigger functions (Stensrud and Fritsch 1994) and convective parameterizations? What is the optimal tradeoff between the number of members and the resolution? How do we synthesize ensemble forecast data so that the operational forecaster need not examine every model solution? These questions and the relevant literature to date are examined in more depth in a workshop summary by Brooks et al. (1995).

A major result of this workshop was the commitment of the National Centers for Environmental Prediction (NCEP) to produce a test set of ensemble for preliminary exploration of the concept of operational short-range ensemble forecasting. A set of 15 ensemble forecasts were produced in each test case, and observations and high-resolution forecasts were also archived. This paper statistically evaluates this set of ensemble forecasts. Three questions are to be answered by this research. First, is the unmodified ensemble useful for evaluating the actual probability of various forecast events, and if not, can this ensemble be postprocessed to generate calibrated probabilistic forecasts? Second, are the assumptions underlying ensemble forecasting met by this ensemble configuration? And third, how do ensemble mean and median forecasts compare to control forecasts from a single, high-resolution model integration?

Section 2 of this paper will review the configuration of the ensemble dataset used in this research. Section 3 evaluates the quality of the ensemble forecasts by an examination of the rank of the verification in comparison to the ensemble, as well as how calibrated probabilistic forecasts can be generated from an imperfect ensemble. Comparisons of ensemble summary statistics to the mesoeta forecasts are provided in section 4. Finally, conclusions and recommendations for future research are provided in section 5.

2. Description of the ensemble and verification data

The test ensemble configuration consisted of 15-member forecasts, 10 of which are integrations of the eta model (Black 1994; Rogers et al. 1995) and 5 from the regional spectral model (RSM) (Juang and Kanamitsu 1994). The eta model ensemble forecasts were run at 80-km resolution with 38 vertical levels out to 48-h lead time. The RSM runs were also run to 48 h at 80-km resolution with 28 sigma levels, the same vertical resolution as the aviation (AVN) run of the Medium-Range Weather Forecast (MRF) Model (Kanamitsu et al. 1991). For comparison, mesoeta forecasts to

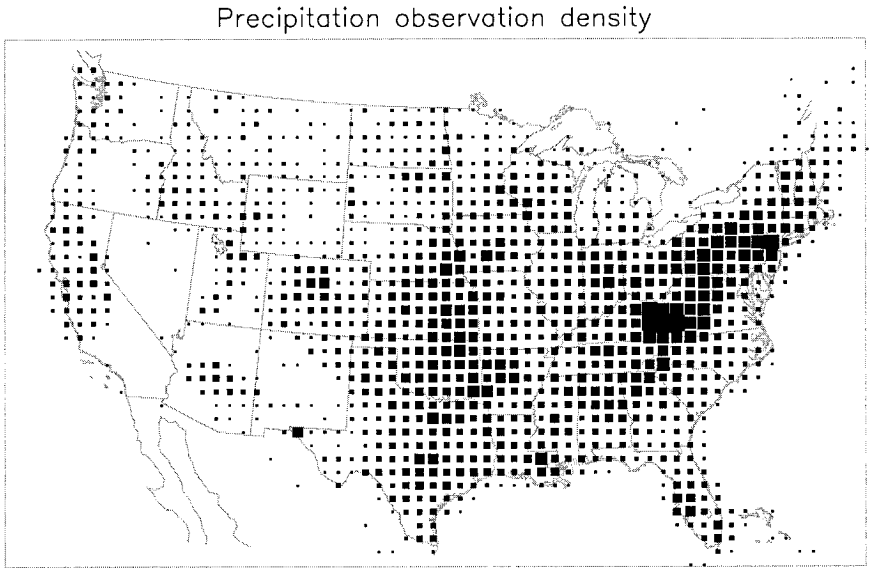

FIG. 1. Map of the location of precipitation observations in the River Forecast Center database. Area of the box is proportional to the number of raw observations assigned to the grid box.

36 h at 29 km and with 50 vertical levels were also archived.

ICs were generated from data already available at NCEP. These included ICs from many of the various in-house analyses and the ICs interpolated from the medium-range breeding forecasts (Toth and Kalnay 1993). For the eta ensemble, ICs were generated from the operational eta IC (Rogers et al. 1996); the eta experimental data assimilation system, or ''EDAS'' (Rogers et al. 1995); the interpolated 3D-variational analysis (Parrish et al. 1996); the Nested Grid Model regional analysis (DiMego et al. 1992); the AVN IC (Parrish and Derber 1992); the MRF model control forecast (Derber et al. 1991; Parrish and Derber 1992); and two positive and two negative bred perturbations (Toth and Kalnay 1993). For the RSM, again the MRF model controls forecast and two positive and two negative bred perturbations were used as ICs.

Boundary conditions for all bred forecasts came directly from the respective bred MRF model forecast. For the EDAS, the previous 12-h AVN forecast boundary conditions were used. For all other forecasts, the on-time AVN boundary conditions are used.

The forecasts were run from 1200 UTC data with the following dates: 5 September, 9 September, 18 September, 25 September, 2 October, 23 October, 8 November, 13 November, 20 November, 27 November, 11 December, 18 December, 26 December 1995; and 23 January and 31 January 1996.

Rawinsonde data within the conterminous United States were used for forecast verification, except for precipitation. For precipitation, 24-h rainfall totals were obtained from the River Forecast Center database. Gridded verification analyses were obtained from the observations by assigning each rainfall observation to its nearest 80-km grid box and then averaging all the observations in each box. The density of observations varies considerably; there are many grid boxes in the intermountain west without any observations; east of the Mississippi, there are usually five or more observations. Figure 1 plots the grid boxes with available observations.

3. Use of the ensemble for generating calibrated probability forecasts

a. Uniformity of verification rank

A number of assumptions are commonly made in ensemble forecasting. First, when one examines an ensemble member's forecast at a given location, the forecast value is assumed to represent an independent sample from the underlying true forecast probability density function (pdf) at that location. Hence, with an infinite set of ensemble forecasts, the relative frequency would converge to the pdf. Also, a perfect forecast model is assumed, and the unknown, true evolution of the atmospheric state is considered a plausible member of the ensemble. Under these assumptions, each forecast should have independent and identically distributed (iid) errors.

These are unrealistically ideal assumptions; any systematic error in the forecast model can result in forecasts with non-iid errors. Similarly, if the ICs are not equally plausible, but some are less likely than others, then the subsequent forecasts cannot be expected to exhibit equal accuracy. Nonetheless, these assumptions permit a way of assessing the quality of the ensemble. Assuming the verification and each forecast are each equally plausible, then the rank of the verification when pooled with N

ensemble forecasts and sorted from lowest to highest value is equally likely to occur in each of the $N + 1$ ranks. Over many independent samples, a distribution of the verification ranks should approximate a discrete uniform distribution with $N + 1$ categories. The hypothesis of uniformity of verification rank can be tested with a chi-square (χ^2) goodness-of-fit test. If the rank distribution is consistently uniform, the ensemble member values may be used to develop calibrated forecast probability distributions of weather events, a major goal of ensemble forecasting.

Rules must be specified for assigning the rank. Matters are simple when the verification is different from all ensemble members. For example, a verification temperature of −3°C when pooled with 10 ensemble forecasts of −5°, −5°, −4°, −1°, 0°, 0°, 1°, 1°, 2°, and 3°C will be assigned rank 4 of 11. For situations where the verification exactly equals some of the forecast members, such as precipitation forecasts of zero and a verification of zero, a new rule for rank assignment was needed. For these cases, the number (M) of members tied with the verification was counted. $M + 1$ uniform random deviates (Press et al. 1992) are generated for the M members and one verification, and the rank of the verification's deviate in the pool of $M + 1$ deviates was determined. All ensemble members with a lower rank had an insignificantly small number (0.0001 in.) subtracted from their values; similarly, all ensemble members with higher rank had the tiny number added. This randomly assigned the rank among the ties without substantially affecting later calculations.

The requirement for independence of the errors at sample points used to populate the rank distribution conflicted with the need for larger sample size. Since the test dataset consisted of ensemble forecasts run roughly once weekly, a sampling of the verification rank at a given location can reasonably be expected to not exhibit significant temporal correlation. Thus, there should be no problem sampling the same locations in each test case. But within one test case domain, how far apart should sample points be in order to be considered independent? Our own examination of the spatial correlation of errors for this dataset (not shown) as well as the results from many other short-range forecast models (e.g., Hollingsworth and Lonnberg 1986; Mitchell et al. 1990; Theibaux et al. 1990; Bartello and Mitchell 1992) suggests a correlation length scale on the order of several hundred kilometers. To boost the sample size, the full rawinsonde data within the conterminous United States was used to populate rank distributions, even though adjacent rawinsonde locations can expect to have error correlations of around 0.4. Similarly, for precipitation, sample points were selected that were 400 km apart from each other, on average. Again, this distance reflects a compromise between some moderate correlation of forecast error between adjacent sample locations and adequate sample size. Specific grid locations to sample were chosen to have at least five precipitation observations assigned to the particular grid box to ensure representativeness of the observation. A map of the sample locations for precipitation and their observation count is shown in Fig. 2.

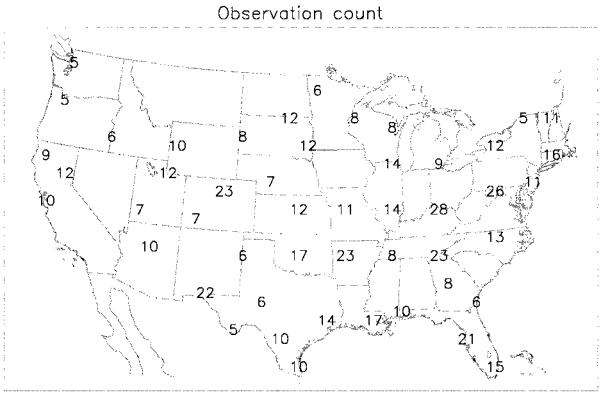

FIG. 2. Observation count at locations for sampling of ensembles for purposes of building rank precipitation rank distributions.

Sampling at these specified locations, distributions of verification ranks were generated. Some representative rank histograms are shown in Fig. 3, here, for 24-h forecasts. The shape of the distributions did not change markedly at other forecast projections (12, 36, and 48 h) or for other fields. Notice the distributions are highly nonuniform; there is a marked tendency for the distributions to be most populated at the extreme ranks. This may indicate systematic errors in the forecast, insufficient variability among ensemble members, incorrect observations, or some combination. If the rank distribution is also skewed, as it is for 500-mb heights in Fig. 3b, this may indicate the systematic bias is large.

The rank distribution is not markedly different in shape when considering subsets of the ensemble members. Figures 4a–c show the rank distributions for the same fields, but here the verification is pooled only with the eta ensemble members initialized from the MRF control and the bred forecasts. Figures 4d–f show rank distributions calculated from the remaining five eta ensemble members. As shown, there is little difference between the bred and nonbred distributions.

Some of the rank distributions, most notably that for the geopotential in Fig. 3b, were highly skewed. To the extent that there are *consistent* systematic errors (biases) affecting these fields, the ensemble may potentially be adjusted. Ideally, it is preferable to archive model forecasts over many seasons and correct for systematic error by location, as are done with Model Output Statistics (MOS; Dallavalle et al. 1992). With the limited set of 15 case days, this is not possible. However, a dataset of this size may be large enough to determine effective, *domain-averaged* corrections for systematic error, where the same corrective adjustment is applied to every member at each sample point. This was attempted here through a cross-validation technique (Wilks 1995), whereby a separate bias correction was determined for

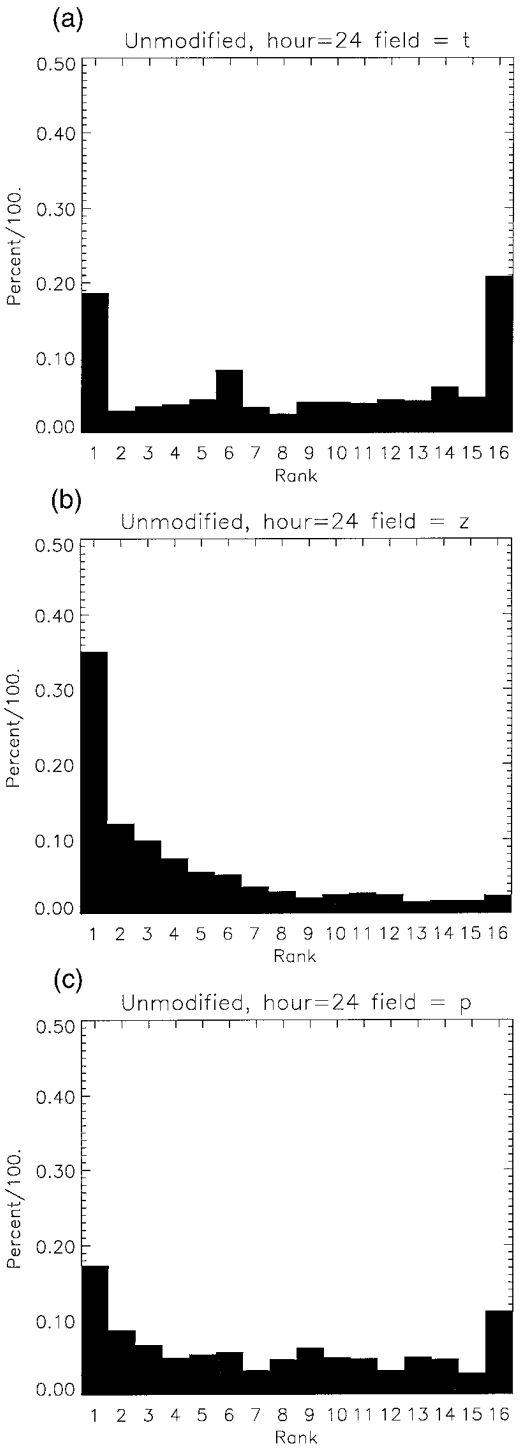

FIG. 3. Rank distributions for 24-h forecasts from the ensemble: (a) 850-mb temperature, (b) 500-mb geopotential height, and (c) 24-h total precipitation amount.

each case day and ensemble member. Each of the other 14 available cases was used to generate the bias correction, and this was applied to forecasts on the remaining day. If the selected bias correction resulted in subzero precipitation forecasts, these were reset to zero temporarily, and then the same correction (0.0001 in.) and methodology explained earlier was either added or subtracted from these member forecasts to set ranks in the case of ties. Rank distributions were then recalculated. The resulting adjusted rank histograms are shown in Figs. 5a–c. Comparing with the unmodified histograms in Fig. 3, the beneficial effect of the bias correction can be most clearly observed with the rank distribution for 500-mb heights (Figs. 3b and 5b), for which the skewness of the distribution is nearly eliminated.

However, after bias corrections, the distributions are each still concave and fail goodness-of-fit tests ($p \ll 0.01$), indicating the hypothesis of uniformity of rank may be rejected. There are potentially many causes; the bias corrections may need to be more sophisticated as with MOS, or the ensemble may truly be insufficiently variable, due to either less than optimal choices for ICs, or model errors. A third possibility is that the observations are not representative of the grid box average and tends to result in extreme rank assignment. For example, with precipitation, the verification is simply the average of all observations within the grid box. There may be situations where a sparse sample of these precipitation observations can result in an estimated grid box average that is not truly representative of the actual average. This is likely to be the case if the precipitation varies substantially within the grid box, and the observation(s) samples its extremes.

To enable us to examine whether observational nonrepresentativeness was a major contributor to the observed nonuniformity, rank distributions for paired sets of ''adequately'' and ''inadequately'' sampled precipitation forecast points were generated and compared. To generate an inadequately sampled precipitation analysis, at each sample point, one observation was selected and used for the verification rather than the average of all observations within the box (which comprised the adequately sampled analysis). The numbers of observations for the adequately sampled analysis are shown in Fig. 2. The distribution of ranks for the adequately sampled set was previously shown in Fig. 3c. The rank distribution for the ''inadequately'' sampled set is shown in Fig. 6. As shown, the rank distributions are quite similar; there are a small number of points that have moved from the extreme rank to the middle, but the distribution is still so highly nonuniform that we conclude the effect of nonrepresentativeness of the observations was typically not a major cause of rank nonuniformity. Thus, the primary cause is assumed to be deficiencies of the model or suboptimal selection of ICs.

b. Uniformity as a function of baroclinic instability

The breeding method used to generate ICs is expected to produce dispersive forecasts in conditions of baroclinic instability (Toth and Kalnay 1993). Of course, the atmosphere is not always baroclinically unstable, so it

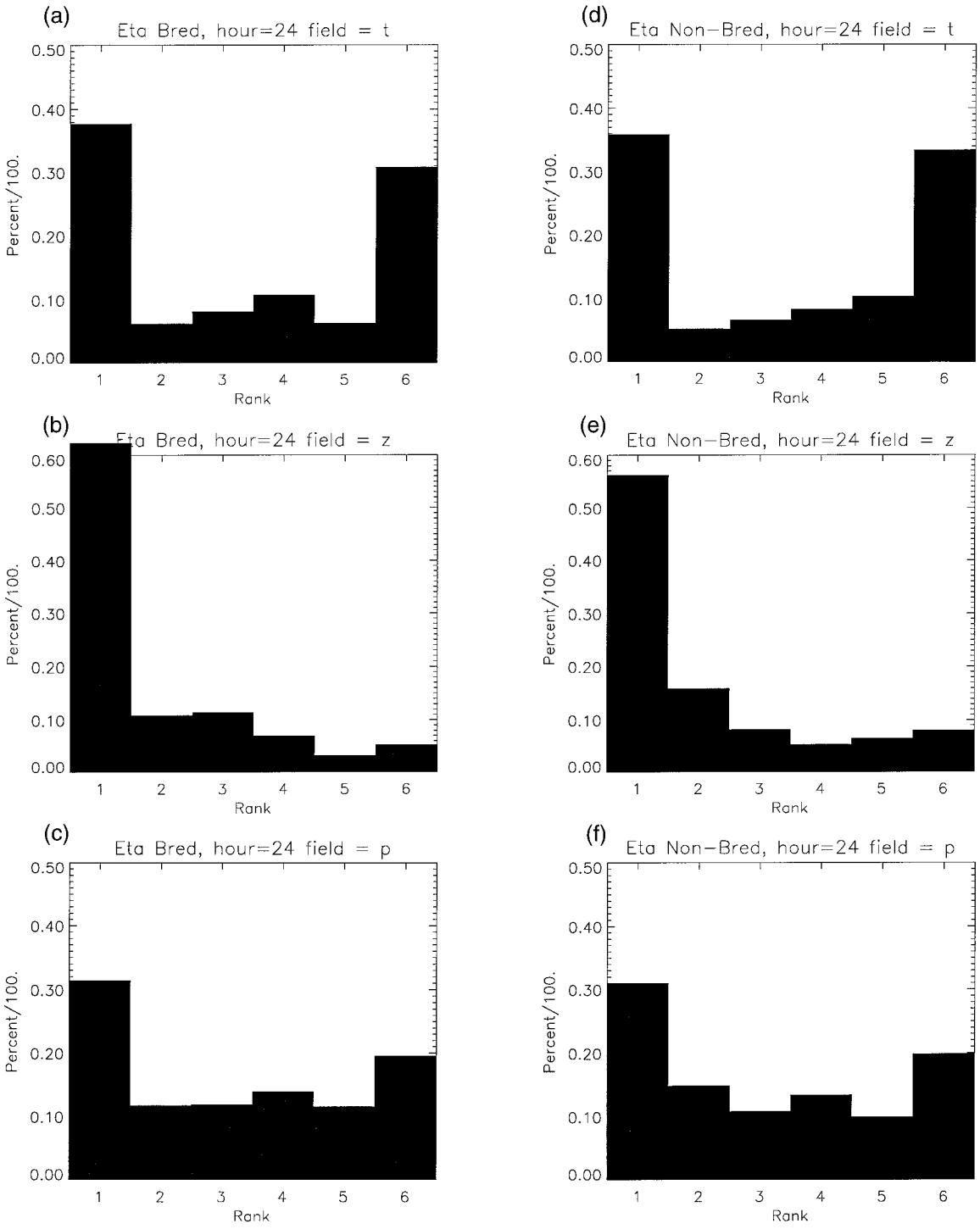

FIG. 4. Rank distributions as in Fig. 3 but now from subsets of eta forecasts from the bred forecast and nonbred forecast initial conditions: (a) 850 temperature from bred, (b) 500-mb geopotential height from bred, (c) 24-h total precipitation from bred, (d) 850 temperature from nonbred, (e) 500-mb geopotential height from nonbred, and (f) 24-h total precipitation from nonbred.

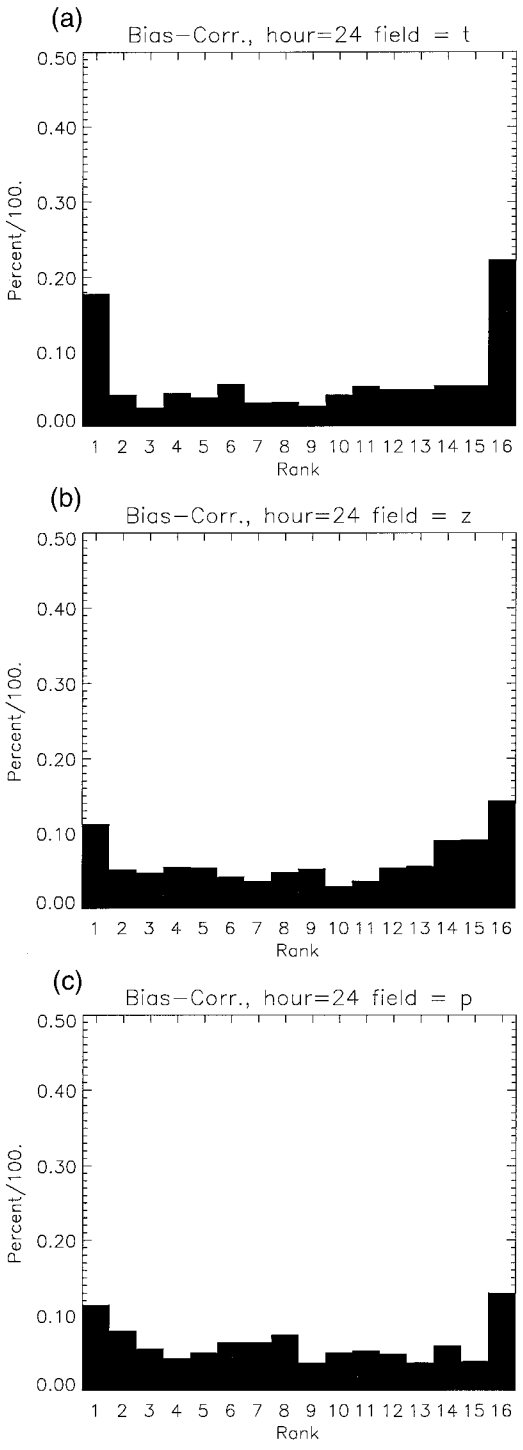

FIG. 5. Same as Fig. 3 but after bias correction: (a) 850-mb temperature, (b) 500-mb geopotential height, and (c) 24-h total precipitation amount.

is reasonable to hypothesize that the dispersion within the ensemble may vary with the weather condition. If so, then the ensemble may be more useful in active weather situations rather than quiescent ones.

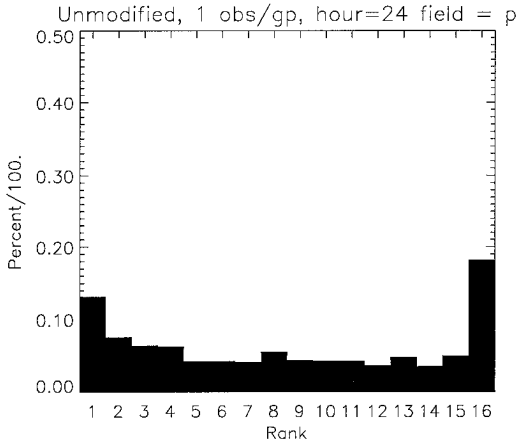

FIG. 6. Rank distribution for 24-h precipitation forecasts with "inadequate" observational sample counts. Compare to Fig. 3c.

To examine this hypothesis, baroclinic instability was measured using 850–700-mb data and the formalism of Lindzen and Farrell (1980):

$$B = 0.3125\frac{f}{N}\left\|\frac{\partial \mathbf{V}}{\partial z}\right\|. \quad (1)$$

Here, f is the Coriolis parameter, N is the Brunt–Väisälä frequency, and $\mathbf{V}$ is the horizontal wind vector. The parameter is larger the more baroclinically unstable the atmosphere. For each sample point where ensemble data were gathered, the value of B at that the forecast time was calculated from the ensemble mean fields. Next, the samples points were sorted into three subsets: the sixth with the lowest values of B, the middle two-thirds, and the upper sixth. The ensemble was corrected for domain-average bias as explained earlier, and rank distributions were then generated for the lowest sixth subset and the highest sixth and compared. For 24-h 850-mb temperatures and 500-mb height forecasts, the plots showed that there was more uniformity of rank for the more baroclinic subset, but not for the precipitation forecasts (Figs. 7a–c). The shape was similar for member forecasts from bred ICs (not shown). Overall, these results suggest that the ensemble may do a better job at predicting the midtropospheric uncertainty under highly baroclinic conditions, but that this does not necessarily translate into improved predictions in the uncertainty of forecasts of surface parameters such as precipitation.

c. Post-hoc corrections to achieve calibration of probabilistic forecasts

Given a nonuniform rank distribution, it is inappropriate to use the ensemble data relative frequencies alone to build probabilistic forecasts. For example, if one-fourth of the ensemble members are above a precipitation threshold, the probability of the event being above the threshold is not necessarily one-fourth, as indicated

by the sum of the highest one-fourth of the ranks. However, there still is useful information in a nonuniform rank distribution; if the shape of the rank distribution remains the same no matter what samples are used to populate it, then it can be used in conjunction with new ensemble data to assess probabilities. For example, consider Fig. 5c, showing a typical rank distribution for bias-corrected 24-h precipitation forecasts. Here the rank distribution indicates that the verification is higher than the highest ensemble forecast on average 14% of the time. Hence, subsequent ensemble forecasts can be sorted, and the highest ensemble member can be used to define the event threshold at which the verification is expected to be greater 14% of the time. Unfortunately, in such a case there is no direct evidence on the distribution of probabilities above the 86th percentile, and the probability of extreme events such as heavy rainfall are of great interest. Hence, an alternative method is necessary to assign probabilities in the tails. With a uniform rank distribution, there is much less probability in the tails, and hence the form of the distribution above the highest ensemble member is of much less concern, especially if there are many ensemble members and hence little probability at the extreme ranks. Hence, though postprocessing in this manner can improve the ensemble post-hoc, it is preferable to correct the model deficiencies that result in the nonuniformity of rank.

We now explore the potential of using rank distributions in conjunction with the ensemble to achieve greater reliability in precipitation forecasts. Suppose there is a sorted ensemble precipitation forecast **X** for a given time and location with N members, a verification observation V, and a corresponding verification rank distribution **R** with $N + 1$ ranks representing the climatological behavior of the verification compared to the ensemble. Then probabilities of forecast events can be assigned using (2):

$$p(V < X_i) = \sum_{j=1}^{i} R_j. \tag{2}$$

The following additional assumptions were also made. First, the rank histogram probability is uniformly distributed below the lowest ensemble member and zero. For a threshold T less than the lowest ensemble forecast X_1,

$$p(0 < V < T) = \left(\frac{T}{X_1}\right)R_1, \qquad 0 < T < X_1. \tag{3}$$

For example, if the lowest ensemble member forecast were 0.03 in., the threshold 0.01 in., and the probability of the verification occurring below the lowest ensemble member 15%, the assigned probability of 0.0–0.01 in. is 5%. Similarly, it is assumed that a given rank's probability is equally distributed between ensemble members:

$$p(X_i < V < T) = \left(\frac{T - X_i}{X_{i+1} - X_i}\right)R_{i+1}, \qquad X_i < T < X_{i+1}. \tag{4}$$

However, assumption of uniformity of probability beyond the highest ensemble forecast X_N is certainly inappropriate. For example, given the highest ensemble forecast is 0.75 in., the probability of 1–2-in. precipitation should be greater than the probability of 2 to 3 in. Hence, it is assumed here that the probability beyond the highest ensemble member has the *shape* of a Gumbel distribution (Wilks 1995) fit to the ensemble data; the Gumbel distribution was chosen for its ability to define rare events in the tails and because of problems defining Gamma distribution parameters over the range of possible precipitation events, especially dry events (Wilks 1990). Given the cumulative distribution function F of the fitted Gumbel distribution, the forecast probability that the verification will occur between any two thresholds $T_2 > T_1 \geq X_N$ is defined as

$$P(T_1 < V < T_2) = \frac{F(T_2) - F(T_1)}{1.0 - F(X_N)}. \tag{5}$$

Probabilistic forecasts of precipitation were generated using this methodology on each of the 15 case days using cross-validation, whereby rank histograms were created from each of the other 14 case days. Because the rank distributions differed in shape with ensemble variability, rather than using one rank distribution to establish probabilities, three separate distributions were used depending on the point's ensemble variability (defined as the standard deviation of the ensemble members about the ensemble mean). The three distributions were for points with ensemble variability below 0.03 in., a second for points between 0.03 and 0.12 in., and another for points with ensemble variability above 0.12 in. Also, because small sample size resulted in some unevenness in the rank histogram, interior ranks were smoothed with a running line smoother (Hastie and Tibshirani 1990) using a neighborhood of 3.0 and a Gaussian kernel with standard deviation of 1.0. The original and resulting smoothed rank histograms used are shown in Figs. 8a–f, here for defining the probabilities for 24-h forecasts from the 5 September 1995 case.

Reliability diagrams (Wilks 1995) were created for the uncorrected and corrected ensemble forecasts [Eqs. (2)–(5) were used to establish probabilities for the uncorrected forecasts, but the rank distribution was assumed uniform]. Figures 9a–c show reliability diagrams for the thresholds 0.01, 0.10, and 0.25 in. (higher thresholds were not shown due to inadequate sample size for high precipitation events). The inset histograms in these figures indicate the relative frequency of usage of each probability category, and a Brier score (Brier 1950) is also indicated. As shown, calibration of the ensemble is greatly improved, with the corrected ensemble in most circumstances being much closer to the desired 45° line indicating perfect calibration; as well, Brier scores are

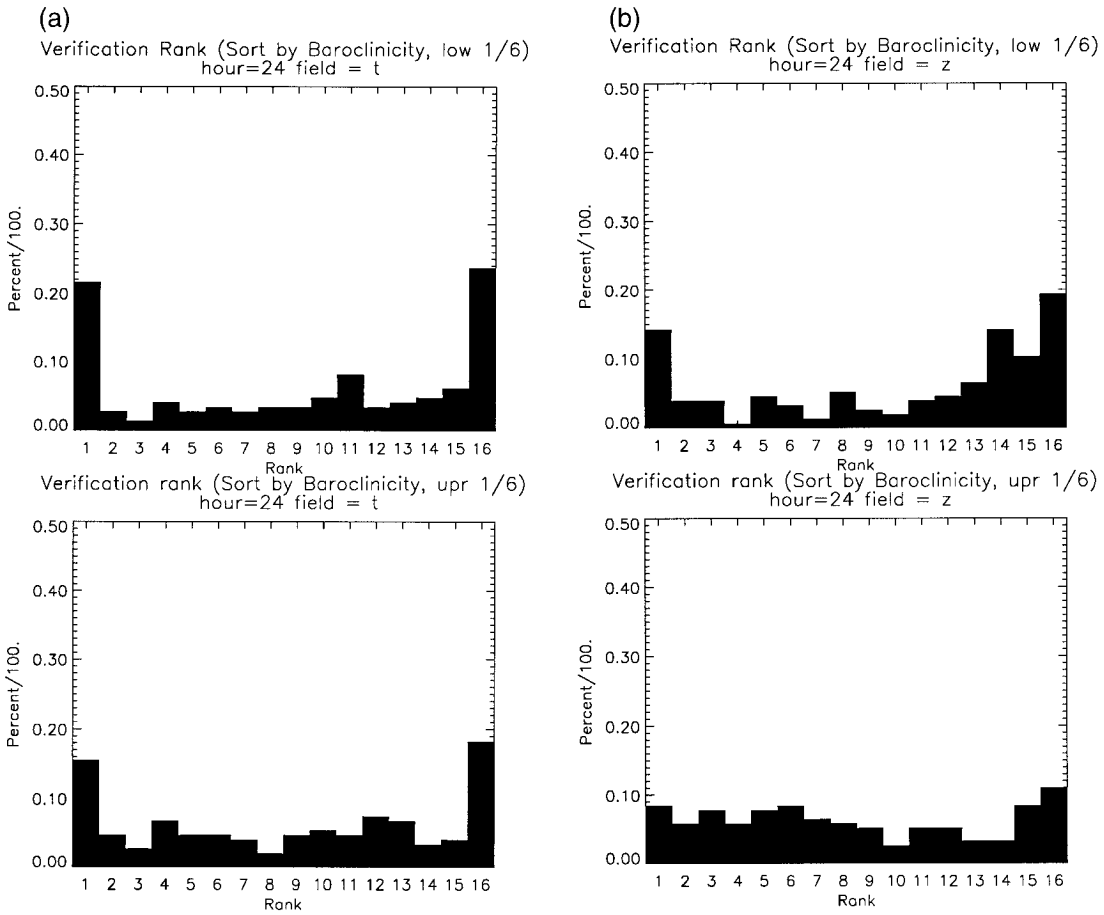

FIG. 7. Rank distributions of 24-h precipitation and 850-mb temperature forecasts for subsets with low and high forecast baroclinic instability: (a) 850-mb temperature, (b) 500-mb heights, (c) precipitation forecasts

improved after the correction. These results suggest that the ensemble may be useful for generating probabilistic forecasts of precipitation. The conclusion section will discuss our future plans for such testing.

d. Errors of each ensemble member

As manifested in the nonuniform rank distributions, the assumptions underlying the production of ideal ensemble forecasts were not met for this dataset. We now check some of these assumptions, specifically here, the assumption that ensemble errors are identically distributed. Tables 1–3 summarize the root-mean-square error (rmse) of each ensemble member after the bias corrections (discussed earlier) were applied. Here, the rmse was calculated for each ensemble member using each sample point on each of the 15 days. Examining the rmse's, it appears there are large differences among members. For example, in Table 1, the RSM 850-mb temperature forecasts, in general, appear to have higher error than the eta forecasts.

To quantify whether the rmse's of the ensemble members differed from each other, resampling tests were performed. Assume there are N_s sample points forecasts over all 15 case days (here, N_s = 883, 929, and 700 sample points for 24-h forecasts of 850-mb temperatures, 500-mb heights, and precipitation, respectively). If a given member's forecast error is identically distributed to errors from the pooled eta ensemble, then the member's rmse statistic calculated from its N_s sample points should be similar to the rmse calculated from a sample of any N_s points from any of the available 10 members times N_s sample points in the eta ensemble, here selected with replacement. Similarly, if the forecast error is identically distributed to the error from pooled eta and rsm forecasts together, as is assumed, the rmse statistic of a given member should be similar to the rmse from a random sample from these pools of 15 members times N_s sample points. Hence, the rmse of each ensemble member was calculated over its N_s sample points. Next, a random set of N_s points selected was with replacement from the pool of $10N$ eta forecasts alone, and the rmse was calculated. This step was repeated 1000 times. Similarly, a set of N_s points was selected from the pool of $15N_s$ combined eta and RSM

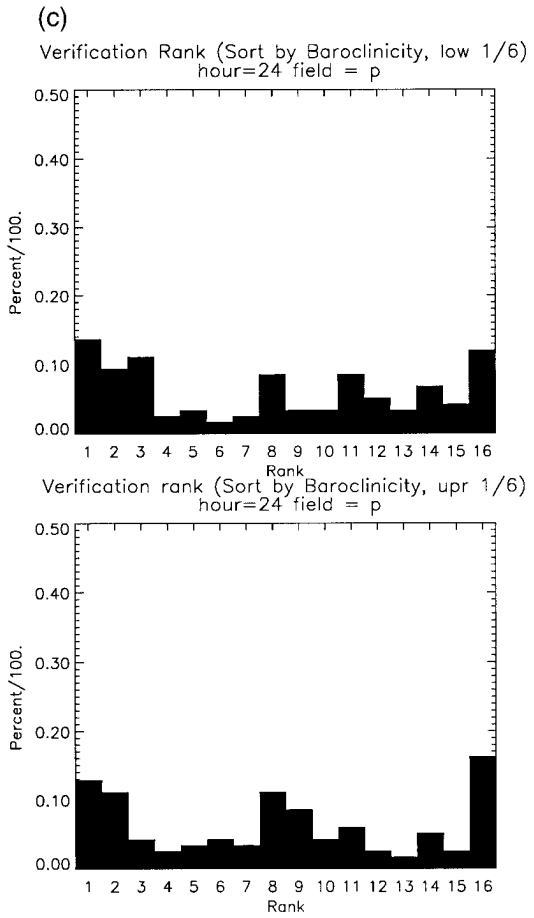

FIG. 7. (*Continued*)

forecasts, the rmse calculated, and the procedure repeated 1000 times. Finally, N_s points were selected from the $5N_s$ RSM forecasts alone, the rmse calculated, and repeated 1000 times.

The rank of the member's actual rmse combined with the 1000 pooled rmse values are shown in columns 3–5 of Tables 1–3. For example, the ranks for 850-mb temperature forecasts in column 3 of Table 1 indicate that RSM forecast members consistently have higher rmse than any random samples calculated from the eta forecasts alone; similarly, column 5 shows that the eta forecasts are typically lower in error than the pool of RSM forecasts. Even considering the eta forecasts alone, it appears several ICs may typically produce lower error forecasts than others. For example, the eta ''control'' forecast has an 850-mb temperature rmse ranked 166th when pooled with 1000 random forecasts, while the eta bred ''N2'' forecasts rank 754th. The variation in performance is more noticeable for 24-h geopotential height forecasts (Table 2) where the ranks of various eta forecasts are consistently near the lowest rank or near the highest rank. However, applying the same resampling with the precipitation forecasts (Table 3), there appears to be greater homogeneity of the rmse, reflected in the typically moderate rank values for each member. Perhaps the more moderate rank is an indication that the assumption of iid error is closer to being met for this forecast parameter.

For geopotential heights and temperature, there is evidence that the forecast errors for each ensemble member are not iid, with the bred forecasts generally having higher error than the nonbred forecasts. Paradoxically, the rank distributions appear similar (Figs. 4a–c vs Figs. 4d–f). The difference might be explained if the bred forecasts were more variable. To test this, the variance of the bred members about their mean were calculated for each sample point, as well as the variance of the five other eta members about their mean. A nonparametric test, the Wilcoxon signed-rank test (Wilks 1995) was used to test the hypothesis of a difference in variances, with the null hypothesis being no difference. The bred members were more clearly more variable ($p < 0.01$ for 850-mb temperatures, $p < 0.00001$ for 500-mb heights). This may imply that dispersion of the ensemble per se is not necessarily a desirable characteristic unless the dispersion is a result of sampling of realistic alternative trajectories through the phase space. This result appears to imply that use of the interpolated bred forecast ICs from the MRF increased the variability but, since the resulting rank distributions were similar, the benefit of increased ensemble variability was offset by decreased accuracy.

4. Comparison of ensemble performance against mesoeta forecasts

Will a single, high-resolution forecast produce a more useful forecast than an ensemble of forecasts run at reduced resolution? With this dataset, forecasts from the 29-km mesoeta forecast model run to 36 h were also available for such comparisons. The processing time of the 10 eta forecasts and the single mesoeta are roughly comparable [31 500 central processing unit seconds (CPUs) for its 48-h forecast vs 26 500 CPUs for the mesoeta's 36-h forecast, respectively].

The usefulness of a forecast is a complicated issue. An ensemble mean forecast may not be valuable for examining the most likely flow pattern, as ensemble averaging can produce unrealistically smooth forecasts. Conversely, the single mesoeta forecast may be less valued by a forecaster interested in the forecast uncertainty. If the forecast user specifically desires the lowest-error forecast at a given point, then the error of the ensemble mean or other summary forecast may provide a competitive alternative to the mesoeta forecast.

The last six rows of Tables 1–3 show the rmse's of the mesoeta forecast and five alternative summary forecasts. Here, the mesoeta forecasts archived on a 40-km grid were area averaged over the appropriate gridpoints to 80-km resolution and verified on the same grid as the ensemble forecasts. Domain-average bias corrections were applied to both the ensemble and mesoeta forecasts by cross-validation, as explained earlier. The

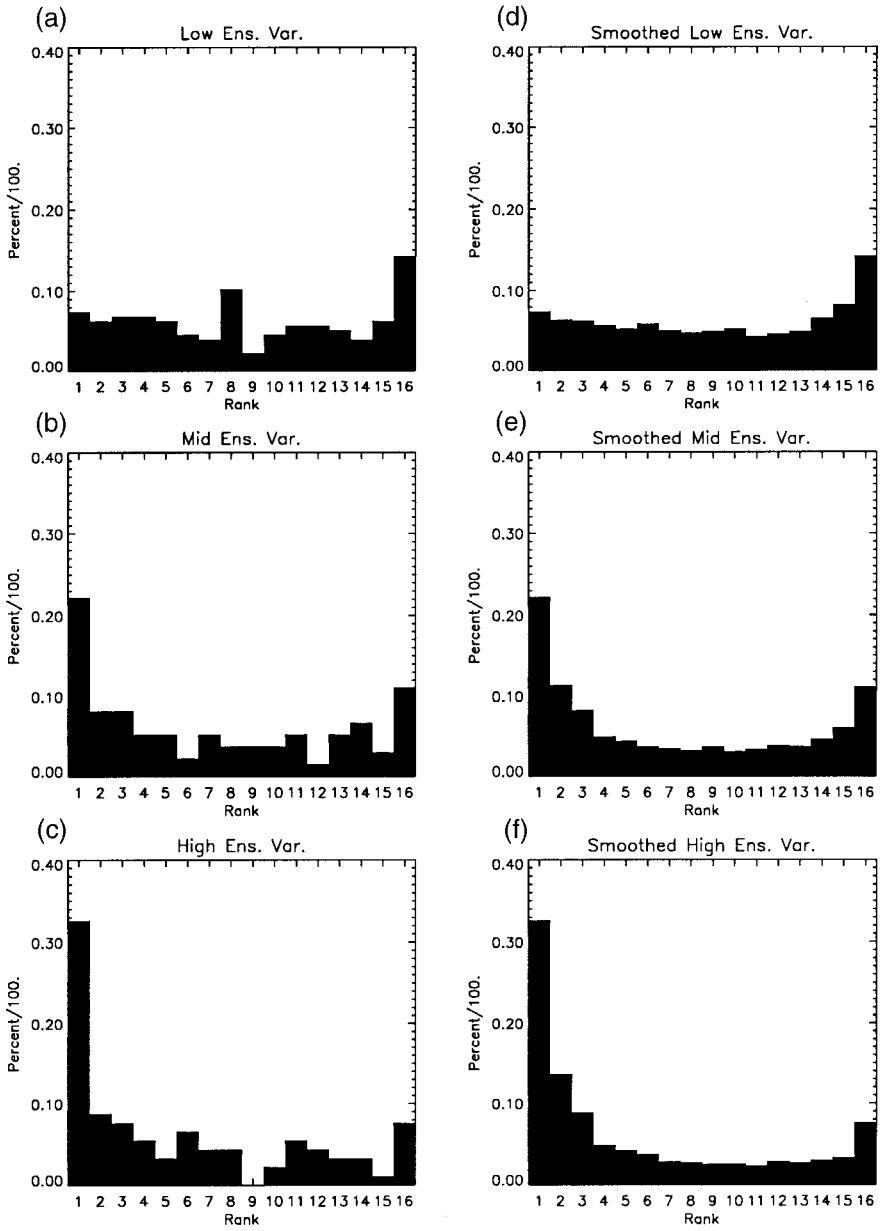

FIG. 8. Rank histograms composited from all casedays except 5 September 1995. (a) Unprocessed histogram for subset of points with ensemble variability less than 0.03 in. (b) As in (a) but for points with ensemble variability between 0.03 and 0.12 in. (c) As in (a) but for points with ensemble variability greater than 0.12. (d) As in (a) but after application of running line smoother. (e) As in (b) but after smoother. (f) As in (c) but after smoother.

five possible ensemble summary forecasts verified here are the mean and median eta forecasts, the median forecast of all 15 members, and two other forecasts, a weighted mean forecast and a forecast of the presumed ''best four'' members. Assuming the error characteristics vary among ensemble members, then perhaps only the best few members or an appropriately weighted sum would minimize the error. Hence, for the weighted mean forecasts, the summary forecast f' was a weighted sum of each of the 15 forecasts:

$$f' = \sum_{j=1}^{15} w_j f_j, \tag{6}$$

with the weight optimally determined using the error variances σ_j^2 of each ensemble member (Daley 1991, 99–100):

$$w_j = \frac{\sigma_j^{-2}}{\sum_{k=1}^{15} \sigma_k^{-2}}. \tag{7}$$

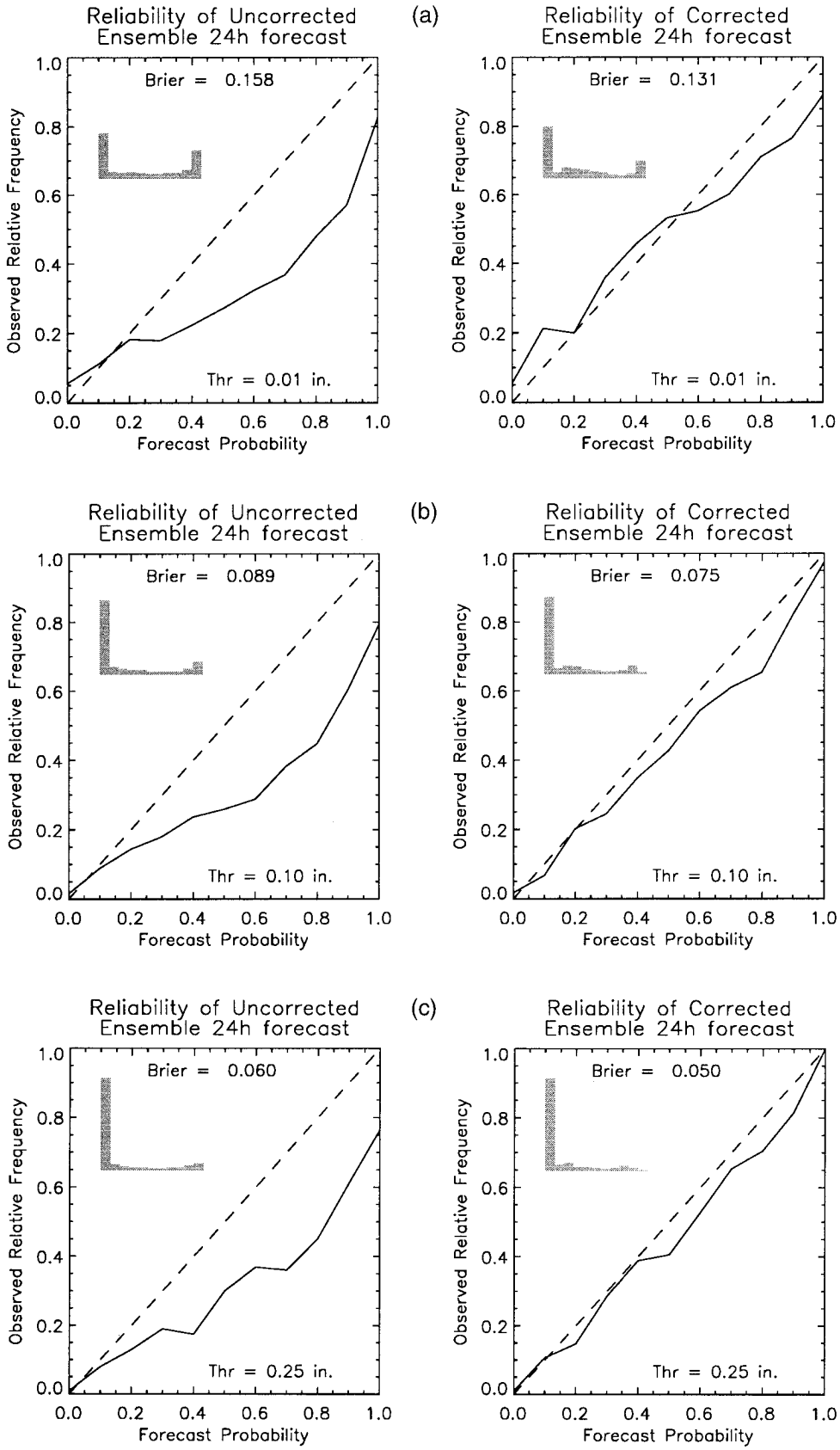

FIG. 9. Reliability diagrams for original ensemble and ensemble corrected with rank histogram information: (a) for 0.01-in. threshold, (b) 0.10 in., (c) 0.25 in.

TABLE 1. Performance of each ensemble member for 850-mb 24-h temperature forecasts. Column 1: Ensemble member. Column 2: The rmse of each ensemble member summed over all points and all case days. Column 3: Rank of the given member's rmse calculated over all points and casedays compared to rmse of 1000 different random shufflings among eta members. Column 4: Ranks of the given member's rmse compared to rmse of 1000 different random shufflings among both eta and RSM members. Column 5: Ranks of the given member's rmse compared to rmse of 1000 different random shufflings among both RSM members alone.

Forecast and initial condition	rms error (°C)	Rank/1001 eta resampled	Rank/1001 all resampled	Rank/1001 RSM resampled
Eta bred P1	2.56	664	241	1
Eta bred P2	2.56	680	246	2
Eta opnl	2.50	443	98	1
Eta AVN	2.44	234	36	1
Eta control	2.42	166	15	1
Eta eDAS	2.48	380	72	1
Eta 3DVAR	2.57	715	277	6
Eta NGM	2.43	216	32	1
Eta bred N1	2.55	647	227	1
Eta bred N2	2.58	754	321	9
RSM control	2.69	946	712	79
RSM N1	2.85	999	968	485
RSM N2	3.03	1001	1000	911
RSM P1	2.98	1001	999	857
RSM P2	2.86	1000	973	528
Mesoeta	2.41	150	14	1
Eta mean	2.40	113	8	1
Weighted mean	2.34	42	6	1
Best four	2.39	110	8	1
Eta median	2.41	154	14	1
Median	2.39	99	7	1

The error variances are again calculated by cross validation, whereby the error variances appropriate to the first caseday are set using a training dataset from casedays 2–15, and so on. Similarly, the ''best four'' is an average of the four ensemble members with the lowest error variances as determined through the cross validation.

Inspecting Table 1, for 850-mb temperature forecasts, it appears the mesoeta forecast rmse is comparable to the values for the individual eta ensemble members but higher than many of the ensemble summary statistics with the exception of the overall median. Conversely, Table 2 indicates that the rmse for 500-mb geopotential height forecasts appears to be better than most of the ensemble summary measures. Finally, in Table 3, the rmse for ensemble summary measures appear slightly smaller than for the mesoeta precipitation forecasts.

These differences were tested for statistical significance using a one-sided nonparametric Wilcoxon signed-rank test with $\alpha = 0.05$ and the null hypothesis being equality of absolute error between the collection of paired samples. Table 4 lists the p values from these tests. As shown, there are differences in comparative performance for the 12-, 24-, and 36-h forecasts. In general, the mesoeta is competitive or better than the ensemble summary measures for 12-h 850-mb temperature, 24-h 500-mb heights, and 36-h heights and temperatures; conversely, for 24-h precipitation, 12-h 500-mb height, and 24-h 850-mb temperature forecasts, all the ensemble summary forecasts are significantly better than the mesoeta. Note also that except for 36-h 500-mb height forecasts, many of the tests for significance are failed in instances where the mesoeta error is lower, indicating the hypothesis of a difference in error

TABLE 2. Same as Table 1 but for 500-mb 24-h geopotential height forecasts.

Forecast and initial condition	rms error (m)	Rank/1001 eta resampled	Rank/1001 all resampled	Rank/1001 RSM resampled
Eta bred P1	30.22	938	829	370
Eta bred P2	29.23	840	620	151
Eta opnl	25.63	89	12	1
Eta AVN	25.49	73	9	1
Eta control	25.54	78	11	1
Eta EDAS	25.51	73	10	1
Eta 3DVAR	28.27	646	392	48
Eta NGM	25.86	119	20	1
Eta bred N1	29.62	887	709	231
Eta bred N2	31.78	996	969	746
RSM control	28.37	673	409	53
RSM N1	31.06	980	921	583
RSM N2	31.25	989	939	630
RSM P1	32.29	998	984	838
RSM P2	32.28	998	984	837
Mesoeta	25.03	38	3	1
Eta mean	25.05	37	3	1
Weighted mean	25.21	54	3	1
Best four	25.04	38	3	1
Eta median	25.15	48	3	1
Median	25.23	54	3	1

TABLE 3. Same as Table 1 but for 500-mb 24-h total precipitation forecasts.

Forecast and initial condition	rms error (in.)	Rank/1001 eta resampled	Rank/1001 all resampled	Rank/1001 RSM resampled
Eta bred P1	0.23	407	441	470
Eta bred P2	0.24	542	600	594
Eta opnl	0.24	534	589	581
Eta AVN	0.24	583	628	625
Eta control	0.24	535	590	582
Eta EDAS	0.24	451	501	513
Eta 3DVAR	0.24	492	545	549
Eta NGM	0.25	610	648	650
Eta bred N1	0.26	732	757	783
Eta bred N2	0.24	542	598	589
RSM control	0.23	386	421	452
RSM N1	0.24	521	574	570
RSM N2	0.23	386	421	452
RSM P1	0.24	485	537	542
RSM P2	0.25	680	703	719
Mesoeta	0.23	373	409	440
Eta mean	0.23	340	374	404
Weighted mean	0.22	206	242	291
Best four	0.25	644	681	697
Eta median	0.23	373	409	440
Median	0.22	239	266	317

TABLE 4. Significance (p values) for one-sided, Wilcoxon signed-rank test of the difference in means between the mesoeta forecasts and various ensemble summary forecasts. An asterisk indicates the overall mesoeta rmse was lower than the given summary measure.

	850 temp	500 height	Precip.
12-h Mesoeta vs eta mean	0.1901*	0.0001	N/A
12-h Mesoeta vs weighted mean	0.0002	0.0000	N/A
12-h Mesoeta vs best four	0.1890*	0.0196	N/A
12-h Mesoeta vs eta median	0.0000*	0.0206	N/A
12-h Mesoeta vs ensemble median	0.1889*	0.0003	N/A
24-h Mesoeta vs eta mean	0.0545	0.0920*	0.0528
24-h Mesoeta vs weighted mean	0.0006	0.2873*	0.4635
24-h Mesoeta vs best four	0.0707	0.3045*	0.2889*
24-h Mesoeta vs eta median	0.0001*	0.0915*	0.1953*
24-h Mesoeta vs ensemble median	0.0182	0.0885*	0.4244
36-h Mesoeta vs eta mean	0.0006*	0.0000*	N/A
36-h Mesoeta vs weighted mean	0.2349*	0.0000*	N/A
36-h Mesoeta vs best four	0.0005*	0.0000*	N/A
36-h Mesoeta vs eta median	0.0003*	0.0000*	N/A
36-h Mesoeta vs ensemble median	0.0678*	0.0000*	N/A

cannot be rejected. In general, it appears that the ensemble summary measures are competitive with the mesoeta.

5. Conclusions and recommendations

This paper examined the performance of a prototype short-range ensemble forecasting system using the eta and RSM models. This prototype was run by NCEP using in-house objective analyses and interpolated bred ICs from the MRF ensembles. The performance of the ensemble was evaluated in two ways. First, the ensemble was assumed to be run with a perfect model and ICs, which were all equally plausible. Under these assumptions, the verification is itself a plausible member of the ensemble, and examining the forecast value at a given point, the rank of the verification when pooled with the ensemble should be equally likely to occur in any of the possible ranks. Using quasi-independent sample points, histograms of the rank distribution were generated, showing that uniformity of rank was not achieved in the unprocessed ensembles. Rank distributions were also examined for collocated sets of points, one representing an average of many precipitation observations, the other using one observation. The distributions were very similar, indicating that nonuniformity of rank could not be attributed solely to problems with the observations but was more a problem of the model and ICs.

The uniformity of rank was also examined for subsets of the ensemble that were the most and least baroclinically unstable. The rank distributions for precipitation were similar but temperature and geopotential height distributions were noticeably more uniform for the high baroclinic subset. This suggests the ensemble is tuned to capture the variability of midtropospheric flow but that this does not necessarily translate to accurate forecasts of the variability of surface parameters.

The error characteristics of individual ensemble members were examined. It was shown that for 850-mb temperature and 500-mb geopotential height forecasts, there were differences in the performance of ensemble members, both between the eta and RSM and even among ensemble members. Typically, the forecasts from the bred ICs were higher in error than the forecasts from the analyses themselves. Precipitation forecasts had more homogeneous error characteristics.

There were some notable forecast benefits demonstrated with this test ensemble configuration. First, even with a nonuniform distribution, it was demonstrated that the ensemble can be post-processed to produce more calibrated probabilistic forecasts. First, simple, domain-average bias corrections were tried. This ameliorated the bias as manifested in the skewness of rank distributions, but the distributions were still more highly populated at the extreme ranks. More sophisticated bias corrections such as are used in MOS were not tried because of the small sample size. For precipitation, a further correction to the ensemble was attempted, whereby probabilistic forecasts were created by using the ensemble in conjunction with the probability information embedded in the rank histograms. This produced a more highly calibrated forecast. We are currently comparing these forecasts against probabilistic precipitation forecasts generated from MOS. Likely there is room yet for improvement in this technique. As with MOS, the more cases available, the more sophisticated bias corrections and more accurate rank histograms can be formulated, improving the probability forecasts. A training set as large as MOS uses may not be necessary, however, as on each forecast day multiple ensemble member forecasts are available, as opposed to a single forecast on each day in the MOS training dataset.

Summary measures such as the mean and median forecasts were shown to often exhibit less error than the competing mesoeta forecasts, especially during the earlier hours of the forecast. This is particularly promising since only readily already available ICs were used; with more carefully selected ICs, it is reasonable to expect improvements in the performance of future SREFs. Further, both our results and the results of Mullen and Baumhefner (1994) suggest that the ensemble performance may be particularly useful in the more synoptically active situations, where good forecasts are particularly important.

How can this ensemble configuration be improved? As demonstrated by the nonuniformity of rank and the heterogeneity of member errors, the amalgamation here of existing analyses and interpolated bred ICs is not optimal. Hence, first and foremost, future research should address just how to best create a distribution of ICs *tailored to short-range weather forecasts.* Ideally,

the ICs should be plausible but should project on the growing modes of the day. The perturbation techniques may vary; what is a good perturbation for a weakly baroclinic, statically neutral summer atmosphere may not be equally appropriate during statically stable, baroclinically active winter months. Also, with the SREF, the design of initial conditions should perhaps focus on the successful prediction of surface parameters rather than midtropospheric flow. We suggest considering the perturbation of fields such as soil moisture, or the use of variable model physics within the ensemble (Stensrud and Fritsch 1994).

Though unexplored here, changes to the forecast model may reduce ensemble errors. Generally, model improvements should benefit both single-integration and ensemble forecasts. However, using the same model physics for single-integration forecasts and ensembles may not always be optimal. For example, the use of strong diffusion coefficients may be preferable to bound the error in a single-integration forecast, but if the diffusion also radically limits the dispersion of the ensemble, the net effect may be detrimental. Recent research by Houtekamer et al. (1996) suggests that diffusion coefficients may be incorrectly formulated and/or too strong.

Assuming a dispersive yet relatively plausible set of ICs can be generated, a third important area needing research is how to allocate the given computer power that is available. This study made no attempt to determine the optimal use of computer power, that is, how many ensemble members at what resolution is optimal given an upper limit for CPU usage. Hence, the ensemble error can reasonably be expected to improve in the future with a more careful selection of ICs and model resolution/member size. As noted before, the choice 10 years hence may be between a single-integration, 5-km model; an 8-member ensemble at 8.4-km resolution; a 16-member ensemble at 10-km resolution; a 256-member ensemble at 20-km resolution; and so on. In the future, we envision that when new computer resources become available, the power will not automatically be used for increased model resolution but rather will be based on test results of the expected benefit at various resolutions and ensemble sizes.

Acknowledgments. This work represents a portion of the first author's (TMH's) Ph.D. dissertation research under the supervision of the second author (SJC) and was supported by the National Science Foundation Grant ATM-9508645. The authors would like to thank Eric Rogers, Zoltan Toth, and Henry Juang at NCEP for producing the ensemble forecasts and Danny Mitchell at NSSL for their archival, and Dan Wilks, Zoltan Toth, and Eugenia Kalnay, Geoff DiMego, Steve Tracton, Harold Brooks, and an anonymous reviewer for feedback that improved the quality of this manuscript.

REFERENCES

Bartello, P., and H. L. Mitchell, 1992: A continuous three-dimensional model of short-range forecast error covariances. *Tellus,* **44A,** 217–235.

Black, 1994: The new NMC mesoscale Eta model: Description and forecast examples. *Wea. Forecasting,* **9,** 265–278.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probabilities. *Mon. Wea. Rev.,* **78,** 1–3.

Brooks, H. E., and C. A. Doswell, 1993: New technology and numerical weather prediction—A wasted opportunity? *Weather,* **48,** 173–177.

——, ——, and R. A. Maddox, 1992: On the use of mesoscale and cloud-scale models in operational forecasting. *Wea. Forecasting,* **7,** 120–132.

——, M. S. Tracton, D. J. Stensrud, G. DiMego, and Z. Toth, 1995: Short-range ensemble forecasting: Report from a workshop, 25–27 July 1994. *Bull Amer. Meteor. Soc.,* **76,** 1617–1624.

——, D. J. Stensrud, and M. S. Tracton, 1996: Short-range ensemble forecasting pilot project: A status report. Preprints, *15th Conf. on Weather Analysis and Forecasting,* Norfolk, VA, Amer. Meteor. Soc., J39–J40.

Daley, R., 1991: *Atmospheric Data Analysis.* Cambridge University Press, 457 pp.

Dallavalle, J. P., J. S. Jensenius Jr., and S. A. Gilbert, 1992: NGM-based MOS guidance—The FOUS14/FWC message. Technical Procedures Bulletin 408, NOAA/National Weather Service, 9 pp. [Available from NOAA/NWS, Services Development Branch, 1325 East-West Highway, Room 13466, Silver Spring, MD 20910.]

Derber, J. C., D. F. Parrish, and S. J. Lord, 1991: The new global operational analysis system at the National Meteorological Center. *Wea. Forecasting,* **6,** 538–547.

DiMego, G. J., and Coauthors, 1992: Changes to NMC's regional analysis and forecast system. *Wea. Forecasting,* **7,** 185–198.

Epstein, E. N., 1969: Stochastic dynamic prediction. *Tellus,* **21**(7), 739–759.

Hamill, T. M., and S. J. Colucci, 1996: Random and systematic error in NMC's short-range Eta ensembles. Preprints, *13th Conf. on Probability and Statistics in the Atmospheric Sciences,* San Francisco, CA, Amer. Meteor. Soc., 51–56.

Harrison, M. S. J., 1994: Ensembles, higher-resolution models, and future computing power—A personal view. *Weather,* **49,** 398–406.

Hastie, T. J., and R. J. Tibshirani, 1990: *Generalized Additive Models.* Chapman and Hall, 335 pp.

Hollingsworth, A., and P. Lonnberg, 1986: The statistical structure of short-range forecast errors as determined from radiosonde data. Part 1: The wind field. *Tellus,* **38A,** 111–136.

Houtekamer, P. L., L. LeFaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.,* **124,** 1225–1242.

Juang, H.-M. and M. Kanamitsu, 1994: The NMC nested regional spectral model. *Mon. Wea. Rev.,* **122,** 3–26.

Kanamitsu, M., and Coauthors, 1991: Recent changes implemented into the Global Forecast System. *Wea. Forecasting,* **6,** 425–435.

Kuo, Y.-H., and R. J. Reed, 1988: Numerical simulation of an explosively deepening cyclone in the eastern Pacific. *Mon. Wea. Rev.,* **116,** 2081–2105.

Leith, C. E., 1974: Theoretical skill of Monte-Carlo forecasts. *Mon. Wea. Rev.,* **102,** 409–418.

Lindzen, R. S., and B. Farrell, 1980: A simple approximate result for the maximum growth rate of baroclinic instabilities. *J. Atmos. Sci.,* **37,** 1648–1654.

Livingston, R. L., and J. T. Schaefer, 1990: On medium-range model guidance and the 3–5 day extended forecast. *Wea. Forecasting,* **5,** 361–376.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.,* **20,** 131–140.

——, 1969: The predictability of a flow which possesses many scales of motion. *Tellus,* **21,** 289–307.

Manikin, G., 1995: Short range ensemble forecasting. M. S. thesis, University of Illinois/Urbana-Champaign, 100 pp. [Available from Department of Atmospheric Sciences, University of Illinois, 105 S. Gregory Ave., Urbana, IL 61801.]

Mason, I., 1982: On scores for yes/no forecasts. Preprints, *Ninth Conf. on Weather and Forecasting Analysis,* Seattle, WA, Amer. Meteor. Soc., 169–174.

McPherson, R. D., 1991: 2001—An NMC Odyssey. Preprints, *Ninth Conf. on Numerical Weather Prediction,* Denver, CO, Amer. Meteor. Soc., 1–4.

Mitchell, H. L., C. Charette, C. Chouinard, and B. Brasnett, 1990: Revised interpolation statistics for the Canadian assimilation procedure: Their derivation and application. *Mon. Wea. Rev.,* **118,** 1591–1614.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.,* **122,** 73–119.

Mullen, S. L., and D. P. Baumhefner, 1994: Monte Carlo simulations of explosive cyclogenesis. *Mon. Wea. Rev.,* **122,** 1548–1567.

Newell, J. E., and D. G. Deaven, 1981: The LFM-II model—1980. NOAA Tech. Memo. NWS NMC-66, 20 pp. [Available from NOAA/NWS, Services Development Branch, 1325 East–West Highway, Room 13466, Silver Spring, MD 20910.]

Parrish, D. F., and J. C. Derber, 1992: The National Meteorological Center's spectral statistical-interpolation system. *Mon. Wea. Rev.,* **120,** 1747–1763.

——, J. Purser, E. Rogers, and Y. Lin, 1996: The regional 3D-variational analysis for the eta model. Preprints, *11th Conf. on Numerical Weather Prediction,* Norfolk, VA, Amer. Meteor. Soc., 454–455.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992: *Numerical Recipes in Fortran.* 2nd ed. Cambridge University Press, 963 pp.

Reynolds, C. A., P. J. Webster, and E. Kainay, 1994: Random error growth in NMC's global forecasts. *Mon. Wea. Rev.,* **122,** 1281–1305.

Rogers, E., D. G. Deaven, and G. J. DiMego, 1995: The regional analysis system for the operational ''early'' eta model: Original 80-km configuration and recent changes. *Wea. Forecasting,* **10,** 810–825.

——, T. L. Black, D. G. Deaven, and G. J. DiMego, 1996: Changes to operational ''early'' eta analysis forecast system at the National Centers for Environmental Prediction. *Wea. Forecasting,* **11,** 391–413.

Simmons, A. J., R. Mureau, and T. Petroliagis, 1995: Error growth and estimates of predictability from the ECMWF forecasting system. *Quart. J. Roy. Meteor. Soc.,* **121,** 1739–1771.

Stensrud, D. J., and J. M. Fritsch, 1994: Mesoscale convective systems in weakly forced large-scale environments. Part III: Numerical simulations and implications for operational weather forecasting. *Mon. Wea. Rev.,* **122,** 2084–2104.

Theibaux, H. J., L. L. Morone, and R. L. Wobus, 1990: Global forecast error correlation: Part I: Isobaric wind and geopotential. *Mon. Wea. Rev.,* **118,** 2117–2137.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.,* **74,** 2317–2330.

Tracton, M. S., 1990: Predictability and its relationship to scale interaction processes in blocking. *Mon. Wea. Rev.,* **118,** 1666–1695.

——, and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Wea. Forecasting,* **8,** 379–398.

Wilks, D. S., 1990: Maximum likelihood estimation for the gamma distribution using data containing zeros. *J. Climate,* **3,** 1495–1501.

——, 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction.* Academic Press, 467 pp.